Assignment -I: Exploratory Data Analysis (EDA) on Education Dataset

Objective

This assignment aims to help you develop skills in performing a detailed exploratory data analysis (EDA) on an education dataset. You will work with various EDA steps, including data cleaning, feature engineering, outlier detection, and advanced data manipulation using pivot tables and apply functions. The analysis should be done in a Google Colab notebook, and the final submission will be evaluated through a viva.

Dataset Link

https://drive.google.com/file/d/1XnxWbFgcwFmbNY4mQAiTaMqL-aB8YGbl/view?usp=sharing

Dataset Description

Use the provided education dataset education_data.csv. The dataset contains the following columns:

- **Student_ID**: Unique identifier for each student (categorical)
- **Age**: Age of the student (numerical)
- **Gender**: Gender of the student (categorical)
- **Department**: Department of study (categorical)
- **Year_of_Study**: Current year of study (categorical)
- **CGPA**: Cumulative Grade Point Average (numerical)
- **Attendance_Percentage**: Percentage of classes attended (numerical)
- **Study_Hours_per_Week**: Hours spent studying per week (numerical)
- **Extracurricular_Activities**: Participation in extracurricular activities (binary: Yes/No)
- **Part_Time_Job**: Whether the student has a part-time job (binary: Yes/No)
- **Satisfaction_Level**: Level of satisfaction with education (categorical)
- **Scholarship**: Whether the student has a scholarship (binary: Yes/No)
- **Parental_Education_Level**: Education level of the student's parents (categorical)
- **Internet_Access**: Whether the student has internet access at home (binary: Yes/No)
- **Family_Income**: Family income level (categorical: Low/Medium/High)

Tasks

1. Data Cleaning (20 points)

Objective: Identify and handle missing values in the dataset. Remove or impute missing values using appropriate techniques (e.g., mean/mode/median imputation for numerical columns, or most frequent category for categorical columns).

Deliverable:

a) Write code to handle missing data, justify your approach, and visualize the changes before and after cleaning.

b) Identify columns with missing values and calculate the percentage of missing data for each.

c) For numerical columns, consider using mean or median imputation depending on the distribution.

d) For categorical columns, consider using the mode (most frequent value).

Visualization: Plot histograms or count plots to show missing data before and after cleaning.

2. Feature Engineering (20 points)

Objective: Create new features that can improve the analysis.

Deliverable: Create at least three new features, describe the rationale behind them, and visualize these new features:

- Create a new feature called `Engagement_Score` by combining `Study_Hours_per_Week` and `Attendance_Percentage`.
- Create a feature `Academic_Standing` that categorizes `CGPA` into 'Low', 'Average', 'High' based on thresholds.
- Derive a feature called `Risk_Level` by combining `Attendance_Percentage`, `Part_Time_Job`, and `Satisfaction_Level`.

Visualization: Use scatter plots, box plots, or bar charts to display the new features.

3. Outlier Detection (20 points)

Objective: Detect outliers in numerical columns, particularly `Age`, `CGPA`, and `Attendance_Percentage`.

Deliverable:

a) Use box plots to visually identify outliers.

b) Calculate z-scores for numerical columns and determine a threshold (e.g., $|z| > 3$) to identify extreme outliers.

c) Handle outliers by capping values or removing rows containing extreme outliers.

Visualization: Plot box plots to show columns with outliers before and after handling them.

4. Advanced Data Manipulation (25 points)

Objective: Perform advanced data manipulation using pivot tables and `apply` functions.

Deliverable:

- Create a pivot table to summarize average `CGPA` by `Department` and `Year_of_Study`.
- Use the `apply` function to create a new feature that indicates whether a student is at high risk based on multiple factors (e.g., Attendance, Part-Time Job, and Family Income).
- Use `groupby` operations to analyze trends in `Engagement_Score` across different `Gender` and `Department`.

Visualization: Use heatmaps or bar charts to visualize insights gained from pivot tables and groupby operations.

5. Plotting Results (15 points)

Objective: Visualize the results of each analysis step effectively.

Detailed Instructions:

- Provide appropriate plots for each task.
- Use histograms for distribution analysis, box plots for outlier detection, scatter plots for feature relationships, and heatmaps for summarizing pivot table data.

Submission Guidelines

- Complete the analysis in a Google Colab notebook.
- Share the Colab notebook link with editing permission restricted to viewing only.
- Ensure that the notebook is well-documented with comments and markdown cells explaining each step.

Deadline: Submission Deadline: **17.12.2024 (fixed, will not change)**

Evaluation Criteria

The evaluation will be conducted through a viva, where you will explain your approach, justify your choices, and discuss the insights gained from the analysis.